

Attention

HE Jiayou

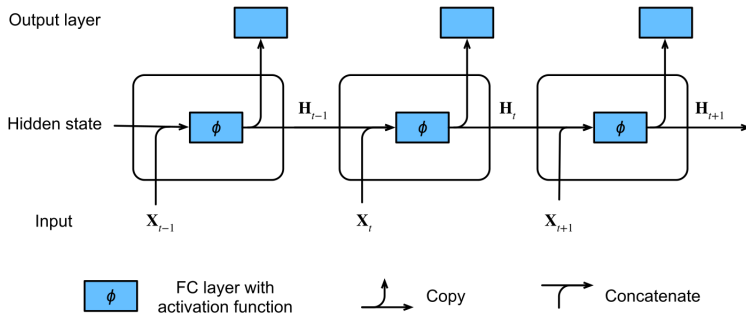
July 6, 2022

Table of Contents

- 1 RNN
- 2 Attention Prompt
- 3 Transformer
- 4 VIT
- 5 Medical-Related
- 6 GNN

Outline

- 1 RNN
- 2 Attention Prompt
- 3 Transformer
- 4 VIT
- 5 Medical-Related
- 6 GNN



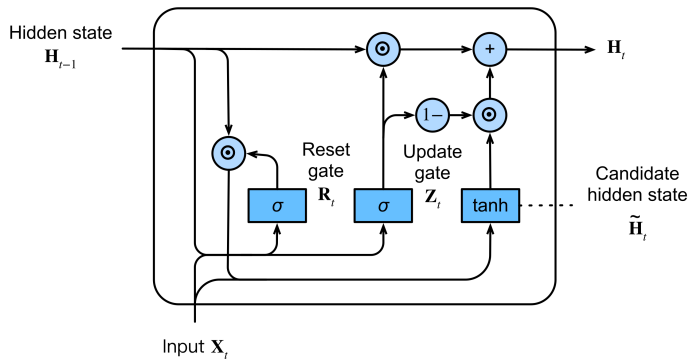
$$H_t = \phi(X_t W_{xh} + H_{t-1} W_{hh} + b_h)$$

$$O_t = H_t W_{hq} + b_q$$

$$X_t \in \mathbb{R}^{n \times d} \quad H_t \in \mathbb{R}^{n \times h}$$

$$O_t \in \mathbb{R}^{n \times q} \quad b_h \in \mathbb{R}^{1 \times h}$$

GRU Gated Recurrent Unit



FC layer with
activation function



Elementwise
operator



Copy



Concatenate

GRU Gated Recurrent Unit

GRU supports gating of the hidden state.

- Reset gates help capture short-term dependencies in sequences.
- Update gates help capture long-term dependencies in sequences.

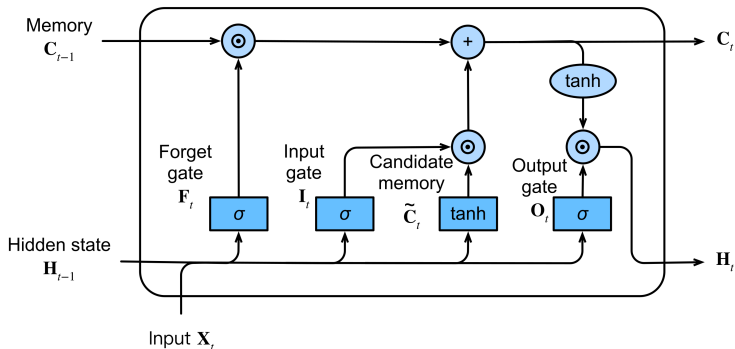
$$R_t = \sigma(X_t W_{xr} + H_{t-1} W_{hr} + b_r)$$

$$Z_t = \sigma(X_t W_{xz} + H_{t-1} W_{hz} + b_z)$$

$$\tilde{H}_t = \tanh(X_t W_{xh} + (R_t \odot H_{t-1}) W_{hh} + b_h)$$

$$H_t = Z_t \odot H_{t-1} + (1 - Z_t) \odot \tilde{H}_t$$

LSTM



FC layer with
activation function



Elementwise
operator



Copy



Concatenate

The idea is similar to GRU.

$$I_t = \sigma(X_t W_{xi} + H_{t-1} W_{hi} + b_i)$$

$$F_t = \sigma(X_t W_{xf} + H_{t-1} W_{hf} + b_f)$$

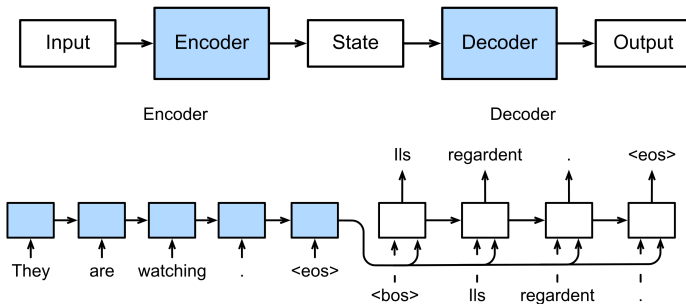
$$O_t = \sigma(X_t W_{xo} + H_{t-1} W_{ho} + b_o)$$

$$\tilde{C}_t = \tanh(X_t W_{xc} + H_{t-1} W_{hc} + b_c)$$

$$C_t = F_t \odot C_{t-1} + I_t \odot \tilde{C}_t$$

$$H_t = O_t \odot \tanh(C_t)$$

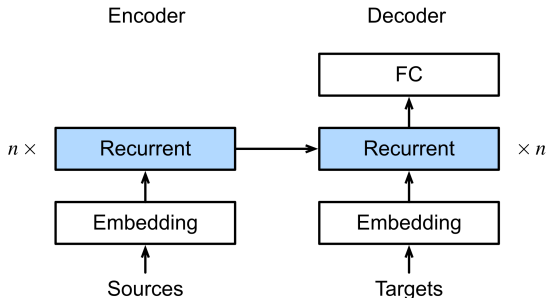
Encoder-Decoder



For the purpose of *variable* input and output sequences.

Encoder-Decoder

- Encoder: $H_t = f(X_t, H_{t-1})$, $C = g(H_1, \dots, H_t)$
- Decoder: to get $P(Y_t | Y_1, \dots, Y_{t-1}, C)$, $H_t = g(Y_{t-1}, C, H_{t-1})$.



Outline

- 1 RNN
- 2 Attention Prompt
- 3 Transformer
- 4 VIT
- 5 Medical-Related
- 6 GNN

Attention Prompt

A simple regression Problem: $f \in \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$.

- Average Pooling:

$$f(x) = \frac{1}{n} \sum_{i=1}^n y_i$$

- Attention Pooling:

$$f(x) = \sum_{i=1}^n \alpha(x, x_i) y_i$$

We call x a *query* and (x_i, y_i) a *key-value* pair.

α is the attention weight, which is the target.

- Nonparametric:

$$\alpha(x, x_i) = \frac{K(x - x_i)}{\sum_j K(x - x_j)}$$

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)$$

- parametric: learnable *Attention Scoring Function*

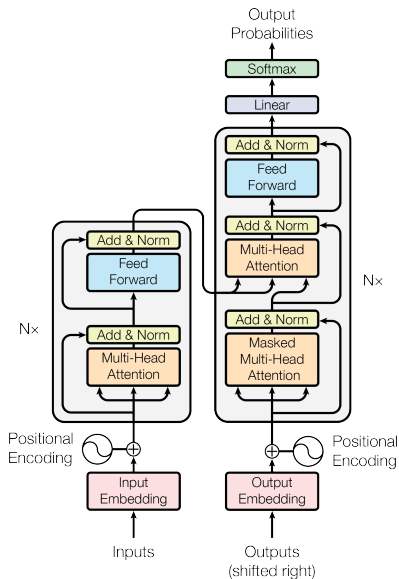
Outline

- 1 RNN
- 2 Attention Prompt
- 3 Transformer**
- 4 VIT
- 5 Medical-Related
- 6 GNN

Transformer

- Relys entirely on self-attention
- Encoder-Decoder architecture
- Positional encoding

Architecture



Encoder:

$N = 6$ layers

Multi-head self-attention +
feed forward

Decoder:

Masked Multi-head
self-attention
Multi-head attention

Others:

Positional Encoding
Layer-normalization

Scoring Function

Scaled Dot-Product Attention

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V$$

$$Q \in \mathbb{R}^{n \times d_k} \quad K \in \mathbb{R}^{m \times d_k} \quad V \in \mathbb{R}^{m \times d_v}$$

Multi-Head Attention

It is beneficial to linear project Q, K, V to d_k, d_k, d_v dimensions h times.

$$\text{Multihead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_n) W^O$$

where $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$

Self-attention:

$$\{f(x_i), 1 \leq i \leq n\}$$

where $f \in \{(x_i, x_i)\}$

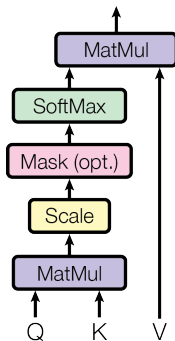
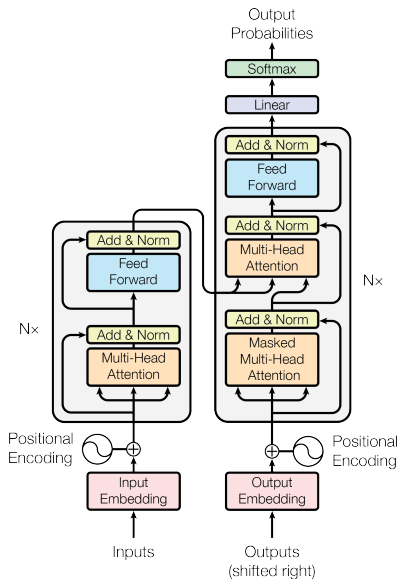
Positional Encoding

For $P \in \mathbb{R}^{n \times d}$:

$$p_{pos,2i} = \sin\left(\frac{i}{10000^{2i/d}}\right)$$
$$p_{pos,2i+1} = \cos\left(\frac{i}{10000^{2i/d}}\right)$$

Make PE be *unique* and *consistent* on space.

Recap



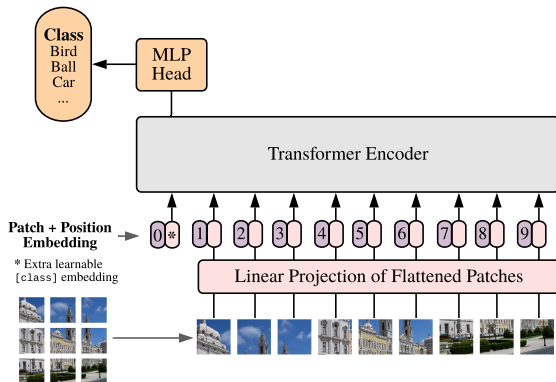
Outline

- 1 RNN
- 2 Attention Prompt
- 3 Transformer
- 4 VIT**
- 5 Medical-Related
- 6 GNN

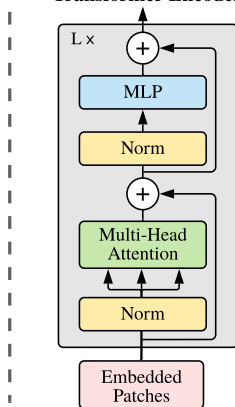
VIT Vision Transformer

Split images into fixed-size patches

Vision Transformer (ViT)



Transformer Encoder



Outline

- 1 RNN
- 2 Attention Prompt
- 3 Transformer
- 4 VIT
- 5 Medical-Related**
- 6 GNN

Attention UNet

Multi-scale Self-guided Attention

Outline

- 1 RNN
- 2 Attention Prompt
- 3 Transformer
- 4 VIT
- 5 Medical-Related
- 6 GNN**

Inductive Bias